DOCUMENT RESUME

ED 466 697 TM 034 276

AUTHOR	Lane, Ken
TITLE	What Is Robust Regression and How Do You Do It?
PUB DATE	2002-02-15
NOTE	16p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, February 14-16, 2002). Some graphs may not reproduce clearly.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Estimation (Mathematics); *Regression (Statistics); *Robustness (Statistics)
IDENTIFIERS	*Outliers

ABSTRACT

All parametric statistical analyses have certain assumptions about the data that must be met reasonably to warrant the use of a given analysis. Distributional normality, for example, is a common assumption. There is a variety of ways that data in a distribution may detract from normality, but one common problem is the presence of outliers. Many applied regression researchers, however, are unfamiliar with the potential role and process of robust regression procedures. Robust regression methods attempt to minimize the impact of outliers on regression estimators, but still invoke parametric assumptions after smoothing the influence of outliers on the slope and intercept. The purpose of this paper is to discuss and demonstrate several robust regression techniques. The paper demonstrates the impact of outliers on regression estimators, discusses several common robust techniques, and illustrates the trimmed least squares and "MM" robust techniques using the S-PLUS statistical software package. A heuristic data set is used to make the discussion concrete and accessible to readers. (Contains 1 table, 8 figures, and 17 references.) (Author/SLD)

What is Robust Regression and How Do You Do It?

Ken Lane

University of North Texas

Paper presented at the al meeting of the annual meeting of the Southwest Edutional Research Association, Austin, TX. February 15, 2002.

Abstract

All parametric statistical analyses have certain assumptions about the data that must be reasonably met to warrant the employment of a given analysis. Distributional normality, for example, is a common assumption. There are a variety of ways that data in a distribution may detract from normality, but one common problem is the presence of outliers. Many applied regression researchers, however, are unfamiliar with the potential role and process of robust regression procedures. Robust regression methods attempt to minimize the impact of outliers on regression estimators but still invoke parametric assumptions after smoothing the influence of outliers on the slope and intercept. The purpose of the present paper is to discuss and demonstrate several robust regression techniques. The paper will (a) demonstrate the impact of outliers on regression estimators, (b) discuss several common robust techniques, and (c) illustrate the trimmed least squares and MM robust techniques using the S-PLUS statistical software package. A heuristic data set will be used to make the discussion concrete and accessible to readers.

What is Robust Regression and How Do You Do It?

The classical least squares estimator is widely used in regression analysis both because of the ease of computation and tradition. Least squares can be traced back to Gauss and Legrendre as far back as 1800 (Wessel, 2000). The rationale behind least squares was to make the residuals very small. Gauss preferred the least squares criterion to other objective functions because in this way the regression coefficients would be computed explicitly from the data. Later, Gauuss introduced the normal distribution for which least squares is optimal.

As time passed it became more apparent that meeting the assumption of a normal error distribution was difficult in data collection. Ordinary Least Squares (OLS) estimation places certain restrictions upon the data in the model. Of interest to this study is the restriction of normal distribution of errors of distribution. This expectation of normalcy is often not attained when studying phenomenon in the real world. Consequently, when OLS is used with data that do not have normal distribution of the errors (i.e., outliers are present), the outlier (s) can significantly influence the estimates. It is well demonstrated that outliers in sample data heavily influence estimates using OLS regression, sometimes even in the presence of one outlier (Nevitt & Tam, 1998).

Data contamination may manifest itself in outliers and other deviations from the standard linear regression model. Outliers could be found along the X or Y-axis and vary in degree in both directions. If an outlier is in the Y direction and has a large residual, it can "potentially influence the regression parameters (i.e., slope and intercept) by pulling the regression line towards the score's Cartesian coordinate so as to minimize the residual error (e) scores" (Serdahl, 1996, p.7). Traditionally the cut off point for outlier is often set at + or – 3 standard deviations from the regression line and having large residuals (sometimes + or – 3 standard deviations from the mean residual of 0). However, as Wiggins (2000) discussed, this is an arbitrary selection and each researcher must take the content under research into consideration when determining what qualifies as an outlier. Because outliers can adversely impact regression results, the purpose of the present paper is to (a) demonstrate the impact of outliers, (b) discuss robust regression as a possible alternative to OLS regression, and (c) provide a brief demonstration of how robust regression can be accomplished.

Effects of outliers

As we will observe even one outlier can skew a distribution. In Figure 1 we see a scatterplot of five data points along a relatively straight line (negative relationship). If one point is miscoded, is a copying error, or is a legitimate outlier and does not fall on the line we see the effect of a single outlier on the least squares regression line. In Figure 2, we see an outlier in the y-direction, and it has a dramatic effect on the OLS line which is now tilted away from the trend of remaining data.

INSERT FIGURE 1 AND 2 ABOUT HERE.

In Figures 3 - 4 we see the effect of an outlier in the x-direction. It has an even more dramatic effect on OLS since it is perpendicular to the actual trend. Because this point has such influence we can denote it as a leverage point. This is because the residual r_i (measured in the y-direction) is enormous with regard to the original OLS fit. A leverage point only refers to its potential for influencing the coefficients. When a point deviates from the linear relation of the majority it is called a regression outlier. Importantly, regression outliers do not always weaken R^2. In Figure 5 we see an outlier along the regression line.

INSERT FIGURE 3 AND 4 ABOUT HERE.

INSERT FIGURE 5 ABOUT HERE.

Even though this point is along the regression line, it is still considered an outlier because it influences (or has leverage over) the strength of the remaining points. It however does not weaken $\underline{R}^2$. Pedhazer (1997) dsscuses how researchers are faced with the delima of what to do with non-normal data. If the researcher deletes the outliers, it should be recorded and reported as such. Pedhazer (1997) encourages performing OLS on the entire data set then repeating the OLS procedure with the non-normal data points removed and reporting both findings. Regardless of what method the researcher eventually applies, to ignore outliers by failing to detect and report outliers is dishonest and misleading (McClelland, 1989, p.231 – 232; Fox, 1991, p.76). Robust regression techniques provide a viable alternative for the astute researcher.

Robust Regression

Modern robust regression techniques, developed mostly during the past 30 years, can provide alternative methods for dealing with nonnormality, and they compete very well with conventional procedures when standard assumptions are met (Wilcox 1998; Wiggins 2000). In response to the impact of outliers upon data when OLS (ordinary least squares) is used several alternative measures have been devised. The alternative measures can be categorized as robust regression measures because they are robust (or resistant) to the outlier impact. Robust estimation methods are considered to perform reasonably well if the errors of prediction have a distribution that is not necessarily normal but "close" to normal (Birkes & Dodge, 1993). Many researchers believe that robust regression is merely dismissing the outliers then performing OLS regression on the remaining subjects. This is untrue as robust techniques act as downweights for the outlying data points. Nevertheless, robust regression methods certainly do reduce the influence of outliers on thefinal solution. Therefore, these methods can be considered as an alternative bridge between ignoring the outliers and deleting the outliers. They are included in the model, but their impact is minimized.

Several robust regression methods are available. Anderson (2001) provides a review of many of the options and evaluated their efficiency in minimizing outlier influence. Several possibilities are mentioned only briefly here. The Trimmed Least Squares Estimator is computationally similar to the trimmed mean. However, the TLS is computed by deleting cases corresponding to a specified percentage of the largest positive and the largest negative residuals under an initial OLS estimation. After case deletion, OLS estimation is performed on the remaining data to compute the TLS estimates of slope and y-intercept. This process has received some criticism (Beasley, 1998), because discarding data creates a situation where data that are systematically missing can lead to biased estimates.

Winsorized regression is used as method to reduce the effect of Y-outliers in the sample by smoothing the observed Y-data rather than simply deleting outlying cases. Winsorization methods modify extreme Y-values by replacing the observed residual for an extreme score with the next closest (and smaller) residual in the data set, and then computing new Y-values using the formulation for an observed score.

MM estimators were developed in 1985 by Yohai. These estimators are defined in three stages. First, the high breakdown estimate is calculated such as in LMS or LTS. Second, an M-estimate of scale s_n with 50% breakdown is computed on the residuals r_i (0*) from the robust fit. Finally, the MM-estimator 0 is defined as any solution of

$$\sum_{i=1}^{n} \psi\left(r_i(\Theta) / s_n\right) x_i = 0$$

Examples using MM and LTS

To perform a regression analysis using MM and trimmed least squares we will use the data set by Rousseeuw and Leroy (1987) (see Table 1) which tracks the number

INSERT TABLE 1 ABOUT HERE

of international phone calls from Belgium in years 1950 – 1973. In this data set the independent variable is the year the data was collected. The dependent variable is the number of phone calls made from Belgium to the United States measured by increments of ten thousand. By observing the data, a different system of measurement was used in the years 1964 – 1969. During those years the total number of minutes was used as the measure rather than the individual telephone call. The result of the least squares regression is depicted in Figure 6. There is heavy contamination caused by a different measurement system in years 1964 – 1969.Instead of the number of phone calls made, the total number of minutes of these calls were reported.

INSERT FIGURE 6 ABOUT HERE

Because of the influence of this measurement contamination, the researchers may wish to invoke a strategy to minimize the influence of errant points. (Of course, if the variable was measured differently, the points probably should just be deleted, but then that would leave us with no illustration!). To illustrate the possible role of robust procedures with these data, both Least Trimmed Squares and MM robust methods were used. S-PLUS for Windows (2000) was used for the analyses. In S-PLUS, the researcher simply needs to follow the Statistics menu to the Regression option, and then select the regression method of choice.

Figure 7 illustrates the new regression line for the Least Trimmed Squares analysis. Comparing the OLS line in Figure 6, it is clear that the outlying points had less influence on the regression line in the robust methods. Here the line comes closer to representing the correctly measured data and is less representative of the errant data.

Figure 8 illustrates the new regression line for the MM estimation method. Again, this line is better representative of the correct data as compared to the OLS method (see Figure 6) and is also superior to the Least Trimmed Squares method (see Figure 7), assuming that the goal is to represent the correctly measured data to the exclusion of the outliers. This result is similar to the findings of Schumacker, et al. (2002) when they compared three robust regression estimators using a large data set.

INSERT FIGURES 7 – 8 ABOUT HERE

Discussion

From the data provided and the analyses we can see that simple least squares regression is not the optimal choice in all circumstances. By using Rousseeuw and Leroy (1987) "phonecall" data we see that using the least squares regression allows errors in coding or copying to unduly affect the regression trend. This calls for alternative measures of regression, which are resistant to such outlier influence. Of course, there is no single best robust estimation procedure. This choice must be made by the researcher and by the context in which the research is being conducted. Nevertheless, robust methods may be useful in contexts where the researcher does not wish to delete outlying points but does not wish to minimize their influence.

Regardless of the robust method used, it will not replace effective and vigilant data editing on the part of the researcher. Pedhauzur (1997) warns against such actions especially since the chores of data analysis tend to be relegated to research assistants that may detect the outliers (and even apply treatment) without the principal researcher an opportunity to examine the data.

References

Anderson, C.R. (2001). A comparison of five robust regression methods with ordinary least squares: Realative efficiency, bias, and test of the null hypothesis. Unpublished doctoral dissertation. University of North Texas, Denton.

Beasley, T. M. (1998). Think different: Comments on alternative regression procedures. Multiple Linear Regression Viewpoints, 25, 83 – 90.

Birkes, D., & Dodge, Y. (1993). Alternative methods of regression. New York: Wiley.

Fox, J. (1991). Regression diagnostics. Thousand Oaks, CA: Sage.

Fox, J. (1997). Applied regression analysis: Linear models, and related methods. Thousand Oaks, CA: Sage.

Hoaglin, D., Mosteller, F., Tukey, J. (1983). Understanding robust and exploratory data analysis. New York: Wiley.

McClelland, G. & Judd, C. (1993). Statistical difficulties of detecting interactions and moderator effects. Pshchological Bulletin, 114, 376 – 390.

Nevitt, J.,& Tam H. P. (1998). A comparison of robust and nonparametric estimators under the simple linear regression model. Multiple Linear Regression Viewpoints, 25, 54 – 69.

Pedhazur, E.J. (1997). Multiple regression in behavioral research: Explaination and prediction (3rd ed). Fort Worth: Harcourt Brace.

Rousseeuw, P. & Leroy, A. (1987). Robust regression and outlier detection. New York: Wiley.

Serdahl. E. (1996, January). An introduction to graphical analysis of residual scores and outlier detection in bivariate least squares regression analysis. Paper presented at the annual meeting of the Southwest Educational research Association, New Orleans, LA. (ERIC Document reproduction Service No. ED 395 949)

S-PLUS 6.0. (2000). Mathsoft, Inc.

Schumacker, R., Monahan, M., and Mount, R. (2002). A comparison of OLS to LTS and MM regression in S-Plus. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Wessel, Paul. (2000). Robust Regression. Retrived January 2, 2002 from: www.soest.hawaii.edi/wessell/courses/gg313?DA_book/node82.html.

Wiggins, B.C. (2000, November). Detecting and dealing with outliers in univariate and multivariate contexts. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY. (ERIC Document Reproduction Service No. ED 448 189)

Wilcox. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? American Psychologist, 53, No. 3, 300 – 314.

Yohai, V. J. (1985). High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics, 15, 642 – 656.

Table 1

Data for Heuristic Illustration of Robust Regression.

Year	Number of Calls[a]
1950	.44
1951	.46
1952	.47
1953	.59
1954	.66
1955	.73
1956	.81
1957	.88
1958	1.06
1959	1.20
1960	1.35
1961	1.49
1962	1.61
1963	2.12
1964	11.90
1965	12.40
1966	14.20
1967	15.90
1968	18.20
1969	21.20
1970	4.30
1971	2.40
1972	2.70
1973	2.90

[a] In tens of millions.

Note: Data used from Rousseeuw and Leroy (1987) Belgian Statistical Survey (Published by the Ministry of Economy).

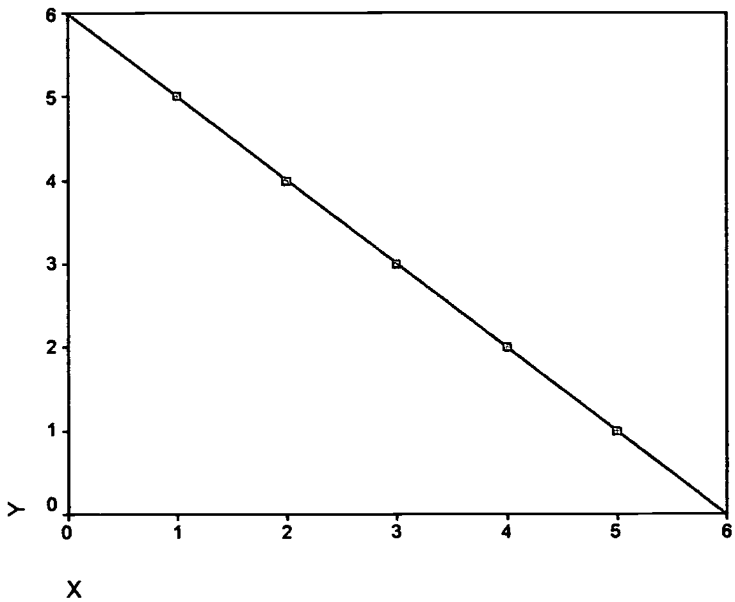

Figure 1. Example of OLS line with no outliers.

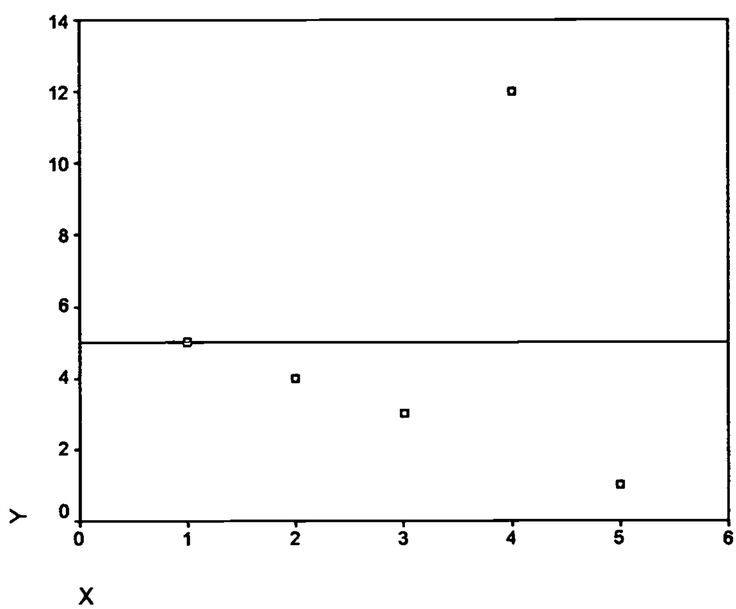

Figure 2. Example of OLS line influenced by an outlier.

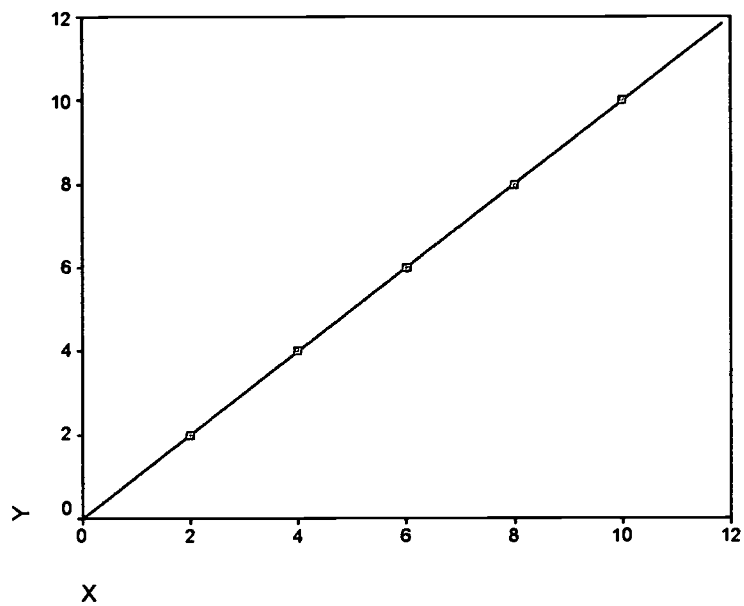

Figure 3. Example of OLS line with no outliers.

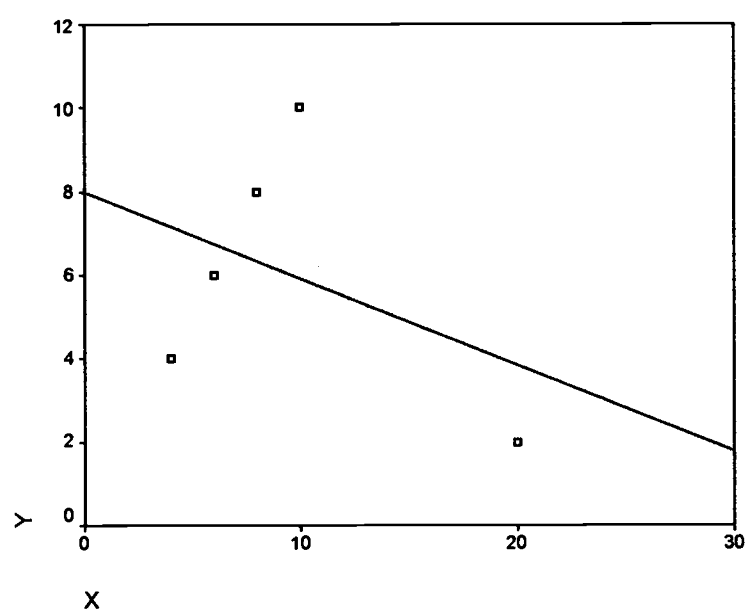

Figure 4. Example of OLS line with and X axis outlier.

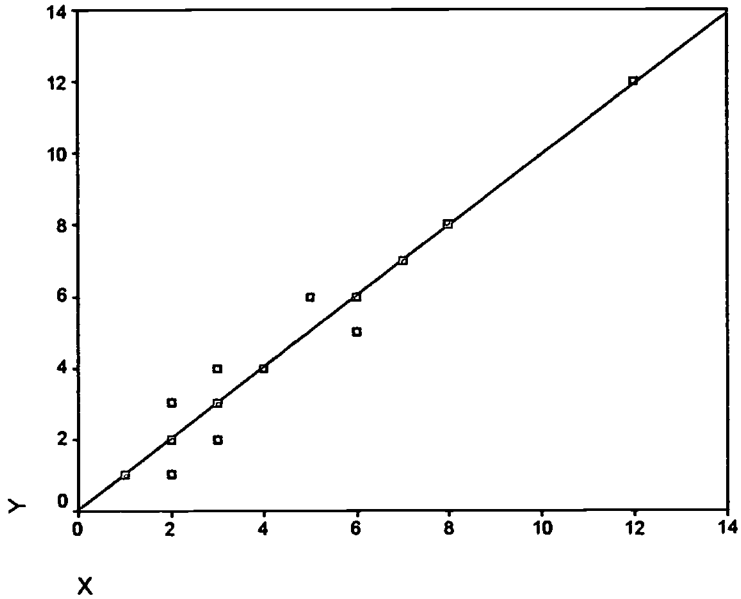

Figure 5. Example of an outlier that does not weaken R^2.

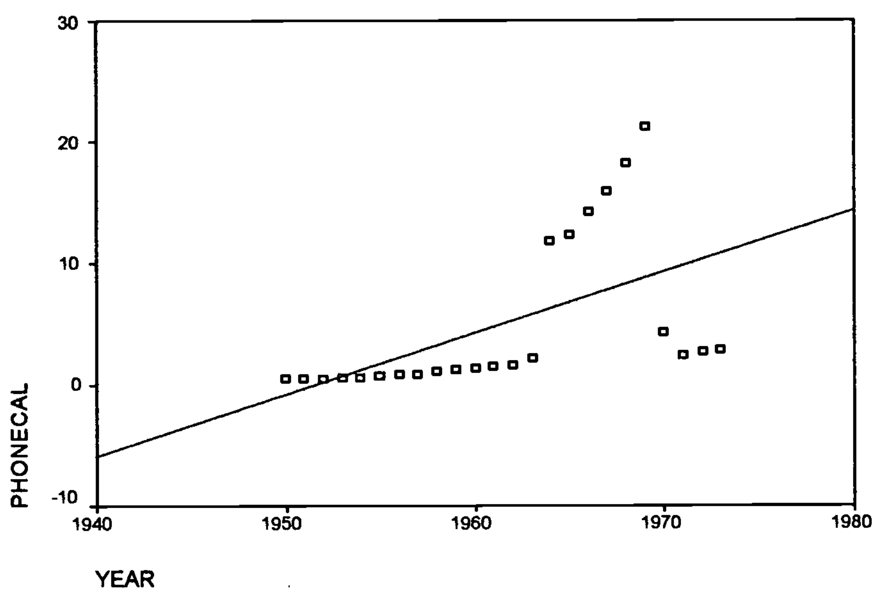

Figure 6. OLS Regression line for example with non-normal error distribution.

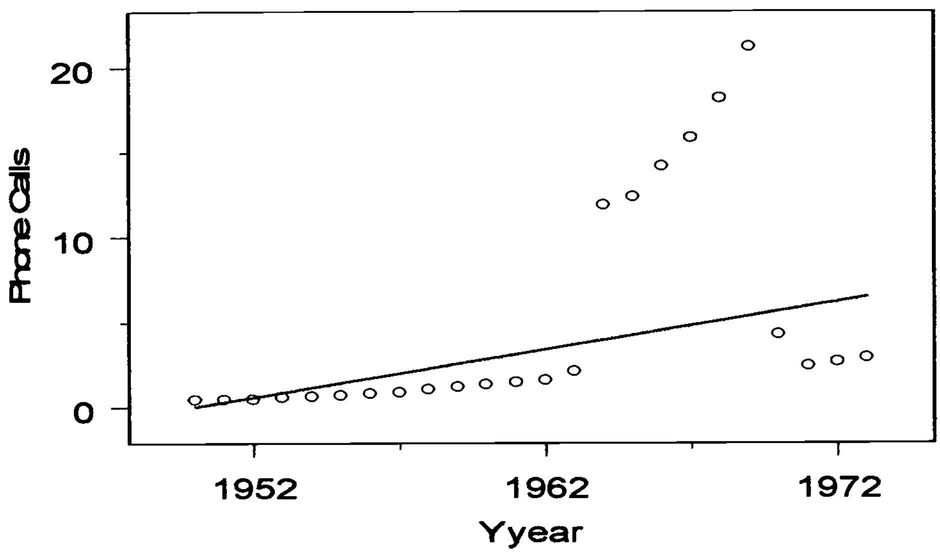

Figure 7. Regression line from Least Trimmed Squares Analysis.

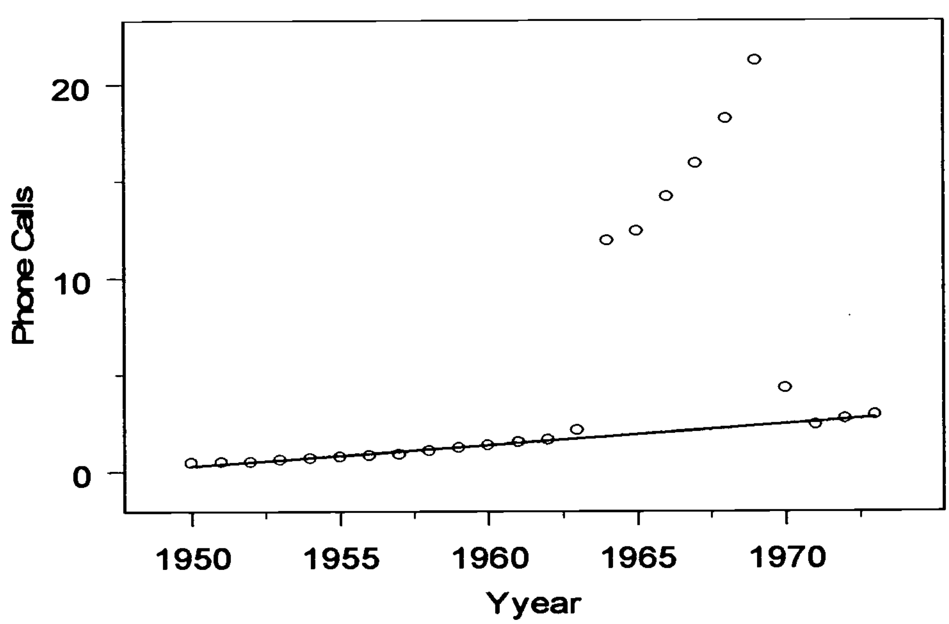

Figure 8. Regression line from MM estimation method.

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: What is Robust Regression and How do you do it?

Author(s): Ken Lane

Corporate Source: University of North Texas

Publication Date: Feb. 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☒	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: Ken Lane / Res. Associate	
Organization/Address: 720 Red Oak Lewisville Tx 75067	Telephone: 972-219-8959	FAX:
	E-Mail Address: laneke@lisd.net	Date: 2/19/02

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com